THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1988 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 263, No. 10, Issue of April 5, pp. 4679-4685, 1988
Printed in U.S.A.

A Structural Domain of the Covalent Polymer Globin Chains of Artemia

INTERPRETATION OF AMINO ACID SEQUENCE DATA*

(Received for publication, September 14, 1987)

Luc Moens‡§, Marie-Louise Van Hauwaert‡, Koen De Smet‡, Daniella Geelen‡, Gonda Verpooten‡, Jozef Van Beeumen¶, Shoshana Wodak‖, Philippe Alard‖, and Clive Trotman‡

From the ‡Department of Biochemistry, University of Antwerp (UIA), B-2610 Wilrijk, Belgium, the ¶Laboratory of Microbiology, State University of Ghent, B-9000 Ghent, Belgium, and the ‖Biological Macromolecular Conformation Unit, Free University of Brussels (ULB) and Plant Genetic Systems, 50 Ave. P. Heger, B-1050 Bruxelles, Belgium

***Artemia* is unusual in having extracellular hemoglobins of M_r 260,000 comprising two globin chains (M_r 130,000), each of which is a polymer of eight covalently linked domains of about M_r 16,000. The amino acid sequence of one of these domains (E1) has been determined. It has 147 residues and M_r of 17,574 including heme. Sequence alignment revealed 19.0% identity with sperm whale myoglobin, whereas other vertebrate and invertebrate globins had between 13 and 24% identity. However, a much higher percentage of residues has a similar side chain character, suggesting that the domain E1 is very similar to other globins in showing the myoglobin fold. Template model building based on the known three-dimensional structure of myoglobin further supports this conclusion. Conversely, the differences between E1 and other globins are believed to reflect differences in the packing of the domains, first in a covalent polymeric subunit containing eight hemes and subsequently by association of two of these subunits as dimers.**

These findings provide further evidence for the versatility of the myoglobin fold.

The primary structure of invertebrate globins is available for only a limited number of species of the phyla Mollusca, Annelida, and Arthropoda. All those sequenced have contained polypeptide chains of M_r about 16,000 binding a single heme group. Alignment of their sequences shows a definite relationship with their vertebrate counterparts (1-4). This has been confirmed by x-ray analysis of *Chironomus, Glycera,* and *Scapharca* hemoglobins which all clearly show the myoglobin fold (5-7).

In contrast, no structural information has previously been available on covalent polymeric globin chains. These proteins with M_r of about 32,000-300,000 and binding 2-16 heme groups are found in the extracellular hemoglobins of some mollusks and arthropods. It is suggested that they are, like the hemocyanins, polymers of structural and functional domains that resemble the low M_r globin chains of the vertebrates (for reviews see Refs. 8-11).

Studies on globin evolution and phylogeny are detailed for vertebrate globins but rudimentary for invertebrate globins (12). Based on protein and gene structure, it is commonly accepted that all globins evolved from an ancestral chain of M_r 16,000 (4, 13). However, the quaternary architecture can differ markedly depending especially on whether the molecule is intracellular or extracellular. Intracellular globins mainly have low M_r, typically 16,000-68,000. In contrast, the extracellular hemoglobins of the invertebrates show a wide variety of M_r values, but their extracellular location makes a high value advantageous in minimizing excretion. A high M_r has been achieved by different routes, exemplified in Annelida by the aggregation of many low M_r chains into a functional hemoglobin (disulfide interactions being involved), or as suggested for some Mollusca and Arthropoda, by concatenation of the low M_r chains into polymeric globins. The covalent polymeric globin chain is, however, a rare departure from the conventional globin, and it was not previously known whether the structure of a polymeric globin could be reconciled with the classical globin model (8-11).

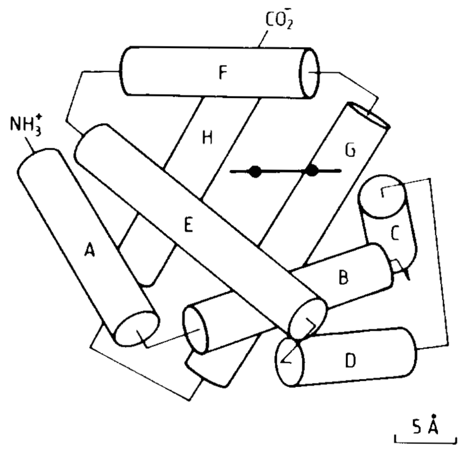
Artemia, a branchiopod crustacean and arthropod, has three extracellular hemoglobins. All three phenotypes have a native M_r of about 260,000 comprising dimers of two globin chains (α and β) of similar size. They bind oxygen with low cooperativity. The $\beta2$ form, which has the highest oxygen affinity, is specifically and reversibly inducible by low environmental oxygen tension. The individual chains have M_r of 130,000 with a minimum M_r of 16,000 calculated from amino acid composition and heme content, suggesting the presence of eight structural and functional units or domains. Cooperativity is not expressed between different domains within the isolated polymeric subunit; it only exists between oxygen binding sites from different globin chains within the native dimer.

Limited proteolysis of the native hemoglobins with subtilisin resulted in a cleavage pattern compatible with a random cleavage of a polymer containing eight domains of M_r about 16,000. A collection of domains (fraction E) was isolated by gel filtration and then further purified to the individual components (E1 to E8) by isoelectric focusing or chromatofocusing (Table 1 (14-20)).

We report below the amino acid sequence of a single domain, E1, isolated from *Artemia* hemoglobin. A combination of sequence alignments and molecular model building shows •that this domain is homologous in primary and probably also in tertiary structure with the low M_r globin chain types.

* This work was supported by the Fund for Joint Basic Research, Belgium (Programs 2.0012.82 and 2.0042.85). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom reprint requests should be addressed.



```
                  10                  20
                  ▾                   ▾
E R V D P I T G L S G L E K N A I L D T W G K V R G
                  |---------------A---------------|

      30                  40                  50
      ▾                   ▾                   ▾
N L Q E V G K A T F G K L F A A H P E Y Q Q M F R F F Q G V Q
|--------------B--------------| |-----C-----|

    60                  70                  80
    ▾                   ▾                   ▾
L A F L V Q S P K F A A H T Q R V V S A L D Q T L L
|----D----| |------------------E------------------|

            90                  100
            ▾                   ▾
A L N R P S D Q F V Y M I K E L G L D H I N R G T
                        |--------F--------|

 110                 120                 130
  ▾                   ▾                   ▾
D R S F V E Y L K E S L G D S V D E F T V Q S F
|-----------------G-----------------|

             140
              ▾
G E V I V N F L N E G L R Q A
|-------------H-------------|
```

FIG. 6. The amino acid sequence of *Artemia* domain E1 shown with a diagram of the three-dimensional structure of *Chironomus* erythrocruorin.

MATERIALS AND METHODS[1]

RESULTS

The amino acid sequence of chain E1 was determined by automated Edman degradation of the amino-terminal segment and by the manual sequencing of peptides obtained by cleavage with trypsin, chymotrypsin, *Staphylococcus aureus V8* protease, and thermolysin. The data relevant in reconstructing the sequence are summarized in Fig. 5.

The majority of residues are well documented, and each is confirmed several times. The exception was that no peptide could be obtained to overlap residues 134 and 135. However, we are confident that there were not any residues missing as the valine at position 135 must have been preceded by a glutamic acid to permit cleavage by *S. aureus V8*; furthermore, the amino acid composition of E1 predicts 147 residues including 21 of glutamic acid, and the final sequence is in agreement (Table 2). A digestion of peptide E1 with carboxypeptidase A indicates an alanine residue at the C terminus. This supports the location of peptide C3 at the C terminus, since alanine is not a chymotryptic cleavage point whereas C3 is generated by chymotrypsin.

The proposed amino acid sequence (147 residues; M_r 17,574 including the heme) is presented along with a globin diagram in Fig. 6.

DISCUSSION

Comparison with all sequences in the NBRF data bank showed our E1 sequence to be most closely related to myoglobins. Many of the highly conserved features characteristic of myoglobins, and familiar in varying degrees throughout the globin family (29), were evident in the sequence of domain E1. As homologous proteins have regions which retain the same general fold and regions where the folds differ (3) we interpreted the sequence of E1 in terms of the "myoglobin fold" and constructed an alignment weighted in terms of the more obviously conserved residues (35) shown in Fig. 7. This alignment revealed that the areas of greatest discrepancy, both of length and of homology, occurred at the loops between the major helices and in the H region. (The standard numbering system based on sperm whale myoglobin is used.) These discrepant regions were investigated by modeling with BRUGEL (30) using a fragment data base approach as described in the Miniprint.

In particular the CD-D and FG regions could both be constructed using fragments from equivalent regions in *Chironomus* globin IIIA whose sequence is also shown in Fig. 7. Homology with the E1 sequence can be traced in globins from diverse phyla, and an alignment with 15 sequences is provided in the Miniprint. *Lumbricus* and *Vitreoscilla* globins were the most difficult to align. Despite the tyrosine at C4 in the authors' (34) alignment of *Vitreoscilla*, which would agree with *Artemia*, we prefer the alignment given in Fig. 8 because it fits better the template of Bashford *et al.* (35) manually applied. To avoid excessive gaps, the overall alignment emphasizes spatial equivalence. Based on this alignment *Artemia* domain E1 revealed 19.0% identity with sperm whale myoglobin, whereas other vertebrate and invertebrate globins had between 13 and 24% (Fig. 8, Table 12).

Region NA (Residues 1-9)—Compared to the vertebrate globins the *Artemia* E1 domain has an amino-terminal extension of 7 residues. Similar extensions are observed in some other globins such as *Anadara*, *Scapharca*, and *Petromyzon* (Fig. 8); however, it is likely that in *Artemia* E1 this segment is part of the linker region connecting two adjacent domains.

Helix A (Residues 10-25)—The high degree of homology with vertebrate myoglobin and hemoglobin chains in the region of residues A1-A16 indicates that this sequence is compatible with an equivalent of the A helix of globins generally. The residue A8, a valine in the myoglobins, is replaced with isoleucine as in some α and β hemoglobin chains preserving a major hydrophobic excursion centered on residues A8-A9. Moreover, the conservation of the key residues A12, A14, A15, and the conservative substitution of A8 (together with G16 leucine to valine where helices A and G would cross)

[1] Portions of this paper (including "Materials and Methods," Tables 1, 2, and 12, and Figs. 1-5, 8, and 11) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are included in the microfilm edition of the Journal that is available from Waverly Press. Tables 3-11 are deposited in the *J. Biol. Chem.* Repository for scientific data. They are available as JBC Document Number 87M-3274B. Orders for supplementary material should specify the title, author(s), and reference to this paper and the JBC document number, the form desired (microfiche or full size photocopy), and the number of copies desired. Orders should be addressed to The Journal of Biological Chemistry, 9650 Rockville Pike, Bethesda, MD 20814 and must be accompanied by remittance to the order of the Journal in the amount of $2.50 for microfiche or $8.00 per set of full-size photocopies. The complete document is included in the microfilm edition of the Journal that is available from Waverly Press.

suggests that the spatial relationship of the A helix to the rest of the molecule is similar to that in the myoglobin fold.

Helical Regions B and C (Residues 27-45)—The key to the assignment of this sequence is the recognition of the critical heme environment which is mainly formed within the sequence B13-B16 and C1-C5. Domain E1 is identical to the myoglobin at positions B13, B14, C1, C2, and C3. All five residues are highly functionally conserved among the vertebrate and invertebrate globins (2, 12), especially the buried hydrophobic residues B13 and B14 and the C2 turn. In contrast residues B15 and B16 are conspicuously variable. C4 which is an almost invariant threonine in all vertebrate globins but rather variable in invertebrate globins (2) is substituted by a tyrosine in E1. Modeling showed this substitution to introduce a number of close atomic contacts which could only be relieved by a concomitant adjustment of the tyrosine side chain and a number of neighboring residues.

It is evident that the region B13-C5 accommodates heme in domain E1 and that the sequence preceding it forms an equivalent of the B helix with glycine at the AB turn. The characters of the well conserved B10 (leucine) and B12 (arginine), the latter preceding the exon boundary in vertebrate globins, are presented by phenylalanine and lysine, respectively. The totality of the sequence in the heme environment supports the identification of the B and C helices in E1. The B6 glycine, characteristic of virtually all globins, permits close crossing of the B and E helices.

Regions CD and D (Residues 50-63)—The CD region displays the highly conserved sequence Phe-*X*-*X*-Phe. Thereafter, the sequence is 1 residue shorter than myoglobin before the recognizable start of the E helix, leading us to question whether a D helix, which is small in myoglobin and nonexistent in α globins, exists. By use of BRUGEL the best match to our sequence from CD1 to E4, of the same length, was found to be the equivalent region of *Chironomus* globin IIIA. The *Chironomus* segment was fitted into myoglobin, and its side chains were changed where necessary to match domain E1. The fit was excellent, showing the E1 sequence to be compatible with a CD structure and D helix of the myoglobin or β globin type, the 1-residue deficiency being taken up in the vicinity of D3 (Fig. 9). The only necessary accommodation was to rotate the side chain of phenylalanine D4, which was otherwise within 0.167 nm of glutamine CD8 (between closest atomic centers). Two instances of structural exchange were striking (Fig. 9): the side chain of the new leucine at D5 took the place of the myoglobin leucine at CD7, and the new D1 leucine similarly replaced D5 methionine.

Helix E (Residues 64-83)—The highly conserved valine at E11, adjacent to the heme, reinforces the assignment of the histidine, 4 residues earlier, as the 6th heme ligand at location E7. On the other hand, the highly conserved E5 lysine and E8 glycine are both substituted by alanine and threonine, respectively; however, these are concerned with a different structural problem of the juxtaposition of helices B and E. Taking E7 and E11 as fixing the orientation of the E helix relative to the heme, it is interesting to find the sequence Ser-Pro-Lys, which is conducive to the start of a helix, then corresponding in location to the Asn-Pro-Lys start of the E helix in β globins. Furthermore, with the exception of E5 and E8 the overall homology with the β globin chains is distinct, confirming the presence of an E helix. The substitutions at

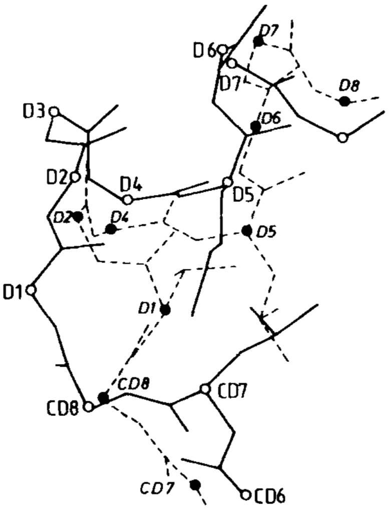


FIG. 9. **The CD and D region.** The CD and D region of *Artemia* domain E1 was modeled with BRUGEL as described under "Materials and Methods." – – –, *Artemia* domain E1; ——, sperm whale myoglobin.

```
                                A      8   12 15      B     6   10         C2  4
MYO WHALE                 VL   SEGEWQLVLHVWAKVE   A   DVAGHGQDILIRLFKS   HPETLEK
ARTEMIA E1         ERVDPITGL   SGLEKNAILDTWGKVR   G   NLQEVGKATFGKLFAA   HPEYQQM
CHIRON III                 L   SADQISTVQASFDKVK   G        DPVGILYAVFKA   DPSIMAK


                   CD 4  7     D            E    5 78 11           EF
MYO WHALE          FDRFKHLK    TEAEMKA      SEDLKKHGVTVLTALGAILK  K KGHHEAE
ARTEMIA E1         FRFFQGVQ    LA FLVQ      SPKFAAHTQRVVSALDQTLL  ALN RFSDQFVYM
CHIRON III         FTQFAGKD    LE SIKG      TAPFETHANRIVGFFSKIIG  ELP NIEAD


                    F  4   8    FG    G                  16    GH  5  H        10
MYO WHALE           LKPLAQSHAT  KHKI  FIKYLEFISEAIIHVLHSR  HPGDF  GADAQGAMNKAL
ARTEMIA E1          IKELGLDHIN   RGT  DRSFVEYLKESLGDSVDEF  TVQSF  GEVIVNFLNEGL
CHIRON III          VNTFVASHKP   RGV  THDQLNNFRAGFVSYMKAH  T  DF  A GAEAAWGATLD


                         22
MYO WHALE          ELFRKDIAAKYKELGYQG
ARTEMIA E1         KQA
CHIRON III         TFF GMIFSKM
```

FIG. 7. **Alignment of the amino acid sequence of *Artemia* domain E1 with the sequences of sperm whale myoglobin and *Chironomus* globin IIIA; the standard numbering system based on myoglobin is used.**

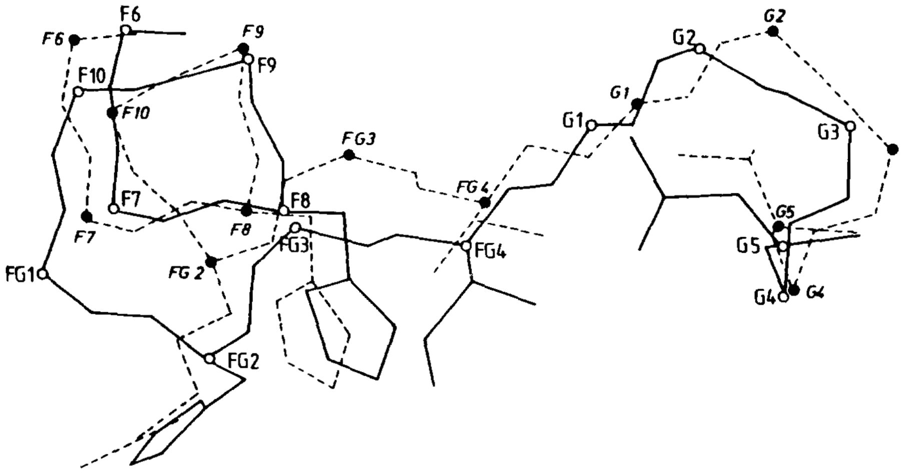


FIG. 10. The FG and G region. The FG and G region of *Artemia* domain E1 was modeled with BRUGEL as described under "Materials and Methods." - - -, *Artemia* domain E1; ——, sperm whale myoglobin.

E5 and E8 could suggest a slightly changed relationship of the B and E helices to each other and may be related to the substitution of the conservative B10 and B12 residues.

EF Turn and F Helix (Residues 84–105)—Reference points locating the putative F helix are the obligatory histidine 5th heme ligand at F8 and the conservative F4 leucine one turn earlier. Having positioned the E and F helices in terms of the two critical sites E7/E11 and F4/F8, the sequence between them is found to be 4 residues longer than in mammalian myoglobins and hemoglobins. The E and F helices are conserved throughout the globins. Thus, the extra residues indicate the possibility of an extended EF turn, but a good match could not be found in the data base. This is one of the most variable regions among globins, *Chironomus* globin IX being longer still (Fig. 8).

FG Turn and G Helix to G6 (Residues 106–114)—Since F and G helices can be identified from the sequence by a combination of structure prediction and alignment, the FG turn can be seen to be 1 residue shorter than myoglobin. By modeling with BRUGEL over the region from F5 to G5, *Chironomus* globin IIIA again provides the best fit, and a striking functional substitution is predicted (Fig. 10). The single turn of π conformation terminating the F helix in myoglobin is substituted with a tighter 3_{10} conformation, thereby omitting FG1 and shifting an arginine into location FG2, where it is positioned to perform the role of H bonding to the heme propionate group previously performed by a histidine in myoglobin.

The threonine assigned to FG4 and the valine at G5 are the only residues in the molecule to contradict a "severe" constraint on side chain character at a buried site in the template of Bashford *et al.* (35). However, examination of the three-dimensional structure with BRUGEL shows these 2 residues to be adjacent, packed against the heme, and also forming a small interior cavity with the side chain of H19. The effect of both replacements is to enlarge slightly this cavity and to introduce some clearance where the heme should be in contact with the protein. However, it is notable that in *Chironomus* globin IIIA the heme is inverted about an axis drawn between the two propionate residues and also tilted to an extent that would displace the edge of its ring about 0.13 nm in the direction of the retracted contact. Examination of all of the contacts between heme and protein in the sequence of domain E1 showed that none is incompatible with the orientation and disposition of the heme being similar to that in *Chironomus* globin IIIA as a model.

G7 to C-terminal (Residues 115–147)—Alignment of the last quarter of the sequence is more problematic because of the absence of strictly conserved residues and meager correspondence to other chains. The latter is characteristic of invertebrate globins (2, 12). It is noteworthy that the sequence from G7 corresponds in vertebrate globins to the exon 3 product (13).

Nevertheless, lengthwise comparison with myoglobin from the start of the G helix places residues of the highest acceptability (zero penalty) in terms of the Bashford (35) template II at G1, G6, G8, G9, G10, G12, G15, G16, and G17, *i.e.* all templated G residues except G5, where the penalty for a leucine to valine substitution is in the lowest category (despite the "severe" constraint). Continuation of this alignment positions a phenylalanine at GH5, where phenylalanine is highly conserved in mammalian and some other globins.

An alternative candidate for GH5 occurs 5 residues earlier if the sequence Ser-Val-Asp-Glu-Phe is advanced into the GH corner. This could result in a more convincing local homology with β globins; however, the consequent shortening of the G helix is not found in any major globin group (35). Furthermore, this alternative phenylalanine is not present in the partially sequenced analogous domain, peptide E7.[2] The compensating increase in H helix that would result is not sought since an exposed H helix adapted to the linking of domains is expected to have been preferentially hydrolyzed, leaving a shortened domain.

Morphological Comparison between Domain E1 and Myoglobin—While clearly the two molecules are superficially related in morphology, the detailed differences reveal adaptations to the requirements of *Artemia.*

An aggregated oxygen carrier has several advantages to an organism without erythrocytes. It minimizes unintentional excretion, provides local high concentrations of active sites with increased thermodynamic efficiency, and facilitates allosteric interaction. In *Artemia* hemoglobin a particle weight of 260,000 is achieved by noncovalent association of M_r 130,000 molecules, each containing eight functional domains, of which peptide E1 substantially comprises one. This domain is not necessarily either the first or last in sequence, and the likelihood is that domain-linking sequences are present at each end but may be incomplete. It will be necessary to assemble the complete sequence of the presumed 8-domain covalent structure before the domain-linking sequences, of which the vestigial H helix may form part, can be identified. Similarly, the 9-residue sequence at the N-terminal of our peptide, the pre-A region, is unfavorable to helical conformation and likely to be assigned a domain-linking function when more sequence is available.

A comparison of the hydropathicity lots of domain E1 and

[2] L. Moens, unpublished data.

myoglobin by the method of Kyte and Doolittle (36) shows regions of similarity and other regions of contrast (Fig. 11). The most striking difference is observed in the CD and D region, where domain E1 is considerably more hydrophobic than myoglobin. In domain E1 the G helix is more hydrophilic and in the early H helix it is more hydrophobic than myoglobin. These differences may reflect an interdomain or intersubunit association in *Artemia* hemoglobin that is not relevant to myoglobin.

At the crossing of the B and E helices a subtle deviation is evident compared with the classic myoglobin structure where B6 glycine and E8 glycine permit close juxtaposition of the two helices; in domain E1 a threonine replaces glycine at E8. Modeling suggests that a rotation of threonine E8 permits close packing of the B and E helices. The retention of many conserved residues in the heme environment, *i.e.* B13–B14, C1–C3, the invariant CD1, CD4, E7, and E11, makes it certain that helices B and E have a myoglobin-like relationship to each other and that a fine difference distinguishes domain E1 from myoglobin at this crossing.

We conclude that the alignment of the sequence of *Artemia* domain E1 with other globins points to the functional conservation of the residues essential to the myoglobin fold. The success of modeling the E1 sequence to the myoglobin template reinforces the conclusion that a myoglobin fold is indeed present in E1. Physical and chemical measurements have shown the *Artemia* hemoglobin to contain subunits of eight domains (Table 1). The ability to translate poly(A^+) RNA and obtain globin chains of M_r 130,000 in reticulocyte lysate or *Xenopus* oocyte (37), together with the low cysteine content of the molecule (14), indicates that the intact subunit is a single translation product rather than linked by disulfide bridges or other post-translational modification. The large (minimum 23 S) size of the mRNA together with the protein structure of *Artemia* globin is suggestive of a gene structure unlike other globins.

This is the first demonstration of the presence of a myoglobin folded structure in a globin of this category.

REFERENCES

1. Dickerson, R. E. & Geis, I. (1983) *Hemoglobin: Structure, Function, Evolution and Pathology*, Benjamin/Cummings, Menlo Park, CA
2. Runnegard, B. (1984) *J. Mol. Evol.* **21,** 33–41
3. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136,** 225–270
4. Hunt, L. T., Hurst-Calderone, S. & Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed) Vol. 5, Suppl. 3, pp. 229–249, National Biomedical Research Foundation, Wash. D. C.
5. Huber, R., Edd, O., Steigemann, W. & Formanek, M. (1971) *Eur. J. Biochem.* **19,** 42–50
6. Padlan, E. A. & Love, W. E. (1968) *Nature* **220,** 376–378
7. Royer, W. E., Jr., Love, W. E. & Fenderson, F. F. (1985) *Nature* **316,** 277–280
8. Wood, E. J. (1980) *Essays Biochem.* **16,** 1–47
9. Chung, M. C. M. & Ellerton, M. D. (1979) *Prog. Biophys. Mol. Biol.* **35,** 53–102
10. Vinogradov, S. (1985) *Comp. Biochem. Physiol.* **82B,** 65–72
11. Van Holde, K. E. & Miller, K. I. (1982) *Q. Rev. Biophys.* **15,** 1–129
12. Goodman, M. (1981) *Prog. Biophys. Mol. Biol.* **37,** 105–164
13. Lewin, R. (1984) *Science* **226,** 328
14. Moens, L. & Kondo, M. (1978) *Eur. J. Biochem.* **82,** 65–72
15. Wood, E. J., Barker, C., Moens, L., Jacob, W., Heip, J. & Kondo, M. (1981) *Biochem. J.* **193,** 353–359
16. D'Hondt, J., Moens, L., Heip, J., D'Hondt, A. & Kondo, M. (1978) *Biochem. J.* **171,** 705–710
17. Moens, L., Geelen, D., Van Hauwaert, M.-L., Wolf, G., Blust, R., Witters, R. & Lontie, R. (1984) *Biochem. J.* **223,** 861–869
18. Moens, L., Van Hauwaert, M.-L. & Wolf, G. (1985) *Biochem. J.* **227,** 917–924
19. Heip, J., Moens, L. & Kondo, M. (1978) *Dev. Biol.* **63,** 247–251
20. Wolf, G., Van Pachtenbeke, M., Moens, L. & Van Hauwaert, M.-L. (1983) *Comp. Biochem. Physiol.* **76B,** 731–736
21. Matsubari, H. & Sasaki, R. M. (1969) *Biochem. Biophys. Res. Commun.* **35,** 175–187
22. Hewick, R. M., Hunkapiller, M. W., Hood, L. E. & Dreyer, W. E. (1982) *J. Biol. Chem.* **256,** 7990–7997
23. Hunkapiller, M. W., Hewick, R. M., Dreyer, W. J. & Hood, L. E. (1983) *Methods Enzymol.* **91,** 399–473
24. Chang, J. Y., Brauer, D. & Wittman-Liebold, B. (1978) *FEBS Lett.* **93,** 205–214
25. Chang, J.-Y. (1981) *Biochem. J.* **199,** 557–564
26. Mendes, E. & Lai, C. Y. (1975) *Anal. Biochem.* **68,** 47–53
27. Rossi-Fanelli, A. & Antonini, E. (1958) *Arch. Biochem. Biophys.* **77,** 478
28. Chen, R. (1976) *Hoppe-Seyler's Z. Physiol. Chem.* **357,** 873–886
29. Takano, T. (1977) *J. Mol. Biol.* **110,** 537–568
30. Delhaise, P., Van Belle, D., Bardiaux, M., Alard, P., Hamers, P., Van Cutsem, E. & Wodak, S. (1985) *J. Mol. Graphics* **3,** 116–119
31. Jones, A. T. & Thirup, S. (1986) *EMBO J.* **5,** 819–822
32. McLachlan, A. D. (1979) *J. Mol. Biol.* **128,** 49–79
33. Chothia, C. & Lesk, A. M. (1986) *EMBO J.* **5,** 823–826
34. Wakabayashi, S., Matsubara, M. & Webster, D. A. (1986) *Nature* **322,** 481–483
35. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 199–216
36. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132
37. Manning, A. M., Ting, G. S., Mansfield, B. C., Trotman, C. N. A., & Tate, W. P. (1986) *Biochem. Int.* **12,** 715–724

Continued on next page.

"A STRUCTURAL DOMAIN OF THE COVALENT-POLYMER GLOBIN CHAINS OF ARTEMIA : INTERPRETATION OF AMINO ACID SEQUENCE DATA"

Luc Moens. Marie-Louise Van Hauwaert, Koen De Smet, Daniella Geelen, Gonda Verpooten, Jos Van Beeumen, Shoshana Wodak, Philippe Alard and Clive Trotman.

MATERIALS AND METHODS

1. Purification of Artemia sp. hemoglobin domain E_1

The Artemia sp. hemoglobin was prepared according to previously published methods (14). The E_1 domain was purified to homogeneity from the subtilisin digested hemoglobin by gel filtration and iso-electric focussing (17-18).

2. Amino acid analysis

An appropriate amount of the apoprotein or peptide was hydrolyzed in evacuated sealed tubes in 6 N HCl for 24 hours at 110°C. The amino acid composition was determined using a Jeol 6AH amino acid analyser.
Protection of tryptophan was performed by the addition of 4% thioglycollic acid during hydrolysis (21).

3. Amino acid sequence determination

Automated amino acid sequence determination was performed using a vapor phase sequencer (Applied Biosystems model 470A) (22).
PTH amino acids were identified and quantified by reverse phase HPLC on a 4,6 mm x 250 mm Cn column (IBM) (22,23).
Manual sequencing was performed using the DABITC-PITC double coupling method (24).
DABTH amino acids were identified either by thin layer chromatography and/or reverse phase HPLC using a Zorbax ODS (4.6 x 250 mm) column (25) DABTH-Leu and DABTH-Ile were identified either by back hydrolysis (26) or reverse phase HPLC using a Zorbax Cn column (25).

4. Heme extraction and denaturation

Heme was removed by acid-acetone precipitation as described (27). After lyophilization, the precipitate was denatured in 6 M guanidine-hydrochloride 50 mM Tris-HCl pH 7.5, 1% 2-mercaptoethanol and finally dialyzed against a volatile buffer.

5. Enzymatic cleavage

Tryptic and chymotryptic digestion was performed in 0.2 M ethylmorpholine-acetate pH 8.3 at an enzyme concentration of 2% during 3 hours at 37°C. Thermolytic digestion was performed in the same buffer at an enzyme concentration of 0.2% during 30 min at 37°C.
Digestion with Staphylococcus aureus V8 protease was performed in 0.1 M ammonium bicarbonate pH 8,3 at an enzyme concentration of 2% during 6-7 hours at 37°C. Cleavage was terminated by lyophilisation.

6. Peptide separation and purification

The acid soluble peptides (0.1% TFA, 5% acetonitrile) were separated by reverse phase HPLC on a 4.6 x 250 mm micro-Bondapack column using a gradient of 0,1% TFA in water to 0.1% TFA in acetonitrile at a flow rate of 1 ml over 120 min. Peptides were detected simultaneously at 220 and 280 nm.
The peptides insoluble in 0.1% TFA, 5% acetonitrile were solubilized by addition of higher % of acetonitrile up to 20%. Separation was as for the soluble peptides but the gradient then started at 20% acetonitrile.
The homogeneity of the peptides was checked by cellulose thin layer chromatography (28).
If necessary, additional purification was performed by preparative TLC.

7. Nomenclature of peptides

Peptides were designated by the following code : T, C, V, Th, tryptic, chymotryptic, Staphylococcus aureus V8 protease and thermolytic peptides respectively. Peptides are numbered for each cleavage according to their position in the amino acid sequence of the chain starting from the amino terminus.

8. Software package

The NBRF-databank was used in combination with software written by P.A. Stockwell, Department of Biochemistry, University of Otago, Dunedin, New Zealand.

9. Model building of the E_1 domain

A 3D model of the E_1 domain was built using the crystal structures of myoglobin (29) as a template and following sequence alignments shown in Fig. 7 and 8. The BRUGEL software package (30) was used throughout this work. Whenever necessary, myoglobin sidechains were replaced by those corresponding to the E_1 sequence. The conformations of the new sidechains were then adjusted to interact optimally with surrounding protein atoms in an automatic procedure which consists in a systematic search for minimum energy conformations as a function of the sidechain dihedral angles.
In both the CD and FG regions the sequence of Artemia E_1 domain is one residue shorter than myoglobin. These deletions were modeled using fragments from equivalent regions in the crystal structure of Chironomus globin III-A (Erythrocruorin). These fragments were chosen as a result of a systematic search through a database containing atomic coordinates and other relevant information on known protein structures, using a procedure similar to that described by Jones (31).
The criteria for the choice were based on structural similarity to the regions in myoglobin immediately preceding and following the deletions and on fragment length - it had to be one residue shorter than in myoglobin. Information on amino acid sequence was not used. The Chironomus fragments were fitted into myoglobin using coordinate superposition (33) and their sequence was changed to match that of the E_1 domain using the procedure for amino acid substitutions described above.

DETAILS OF SEQUENCE DETERMINATION

1. Amino acid composition of chain E_1.

In Table 2 the experimentally determined amino acid composition for E_1 (18) is compared with the number of residues derived from the proposed sequence.

2. Sequence determination

2.1. Automated sequence determination

The amino terminal sequence of chain E_1 was determined up to residue 74 starting from 5 nmol apoprotein as described in materials and methods. Although at each cycle several amino acid residues were identified.the sequence could be deduced unambiguously from their relative variation in concentration taking into account the previous and next cycle (Table 3).

2.2. Manual sequence determination

An appropriate amount (13-15 mg) of E_1 was digested respectively with trypsin , chymotrypsin#, Staphylococcus aureus V8 protease and thermolysin. The resulting acid soluble peptides were separated by reversed phase chromatography (Fig. 1-4). The isolated peptides resulting from the respective cleavage were pure enough, as judged by thin layer chromatography, to determine their amino acid composition (Table 4,6,8,10) and sequence (Table 5.7,9.11).
One peptide, T_{10}, was isolated by preparative thin layer chromatography from the acid insoluble fraction after trypsin digestion.
The manual sequenced peptides covering residues 1-74 completely confirm the results obtained by automated sequence determination (Table 3, Fig. 5).

The data from automated and manual sequencing and the amino acid compositions of the relevant peptides (Fig. 5), summarized in Table 3 to 11 are deposited with the Journal of *Biological Chemistry* who will reproduce them on individual demand.

3. Reconstruction of the sequence of chain E_1

The amino acid sequence of chain E_1 was reconstructed from the above summarized data. Fig. 5.

Table 1 : Characteristics of the extracellular hemoglobins of Artemia sp. (ref 14-20).

	Hemoglobins			Globin chains	Domains
$s_{20,w}$	11.57 ± 0.1			6.17	---
Mr	260,000			α : 131,000 β : 130,700	15,000-17,000
Heme content	1/17,000			1/17,000	1/17,000
Dimensions	120 x 70 Å			---	---
Phenotype	HbI	HbII	HbIII	---	---
Globin composition	α_2	$\alpha\beta$	β_2	---	---
P_{50} O_2 (mm Hg)	5.34	3.70	1.80	3.45	0.36
Hill coefficient	1.62	1.92	1.58	1.00	1.00
pI	5.60	5.70	5.90	---	4.6-5.9

Table 2 : Comparison of the amino acid composition of chain E_1 (ref 18) with the number of residues derived from the proposed sequence.

A.A.	Determined(Ref 18) Residue/mol	Derived from the proposed sequence
Asp	14.43	14
Thr	6.87	7
Ser	7.73	8
Glu	21.09	21
Pro	4.83	4
Gly	10.82	12
Ala	9.92	10
½ Cys	-	-
Val	14.12	13
Met	1.12	2
Ile	5.32	5
Leu	17.72	17
Tyr	3.40	3
Phe	11.10	12
Trp	1.03	1
Lys	6.82	7
His	3.55	3
Arg	8.21	8
	148	147

Table 12 : Percentage of amino acid sequence identity of some globins with sperm whale myoglobin.

Species	Location	Percentage identity
Artemia domain E1	extracellular	19.0
Chironomus thummi IIIA	extracellular	17.0
Chironomus thummi IX	extracellular	17.0
Scapharca inaequivalvis	intracellular	19.6
Aplysia limacina	intracellular	20.9
Anadara broughtonii	intracellular	19.6
Glycera dibranchiata	intracellular	20.9
Tylorrhynchus heterochaetus	extracellular	13.1
Lumbricus terrestris	extracellular	18.3
Vitreoscilla sp.	intracellular	15.0
Lupinus luteus lehgemoglobin	intracellular	17.6

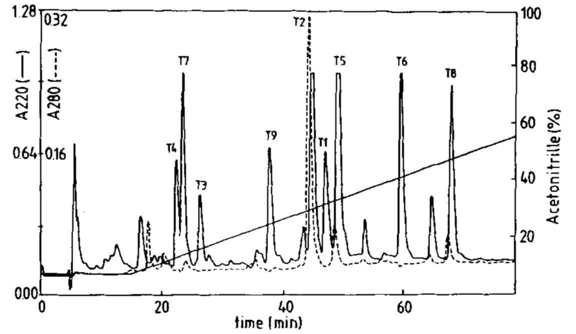


Fig. 1 : Reverse phase chromatography of the tryptic peptides from Artemia domain E_1. Buffers were 0.1% TFA (A) and acetonitrile (B). A linear gradient was developed from 0% A to 100% B in 120 min at a flow rate of 1 ml/min. Peptides were numbered according to their order in the final sequence. Detection was performed at 220 nm (——) and 280 (---).

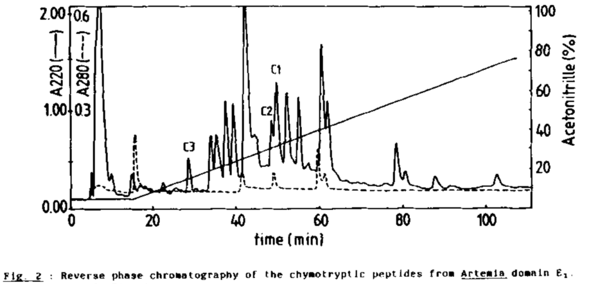


Fig. 2 : Reverse phase chromatography of the chymotryptic peptides from Artemia domain E_1. Conditions as in Fig. 1.

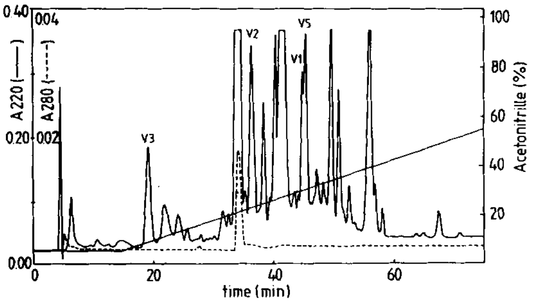


Fig. 3 : Reverse phase chromatography of the Staphylococcus aureus V8 peptides from Artemia domain E₁.
Conditions as in Fig. 1.

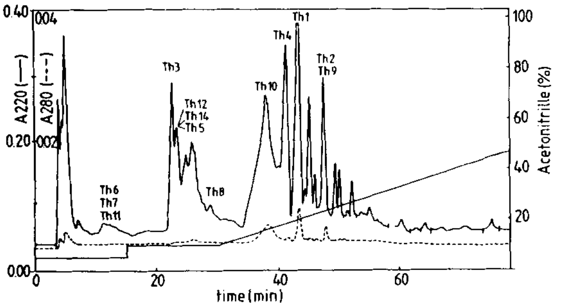


Fig. 4 : Reverse phase chromatography of the thermolytic peptides from Artemia domain E₁.
Conditions as in Fig. 1.

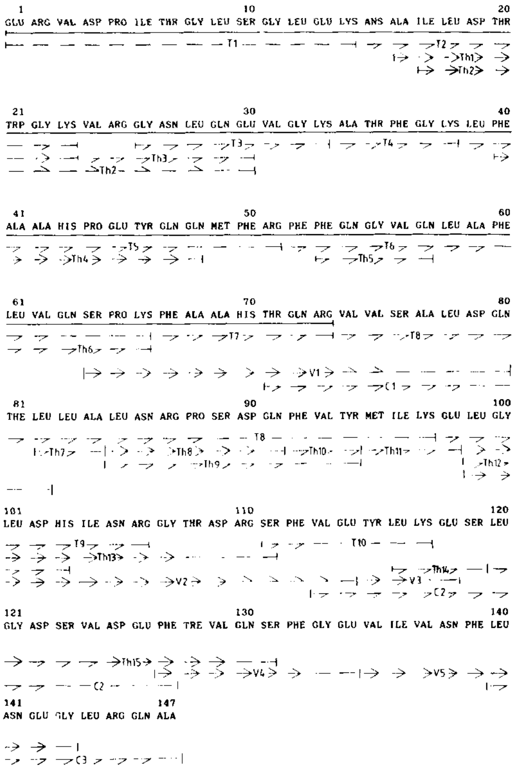

1 10 20
GLU ARG VAL ASP PRO ILE THR GLY LEU SER GLY LEU GLU LYS ASN ALA ILE LEU ASP THR
T1
T2
Th1
Th2
21 30 40
TRP GLY LYS VAL ARG GLY ASN LEU GLN GLU VAL GLY LYS ALA THR PHE GLY LYS LEU PHE
T3
T4
Th3
Th2
41 50 60
ALA ALA HIS PRO GLU TYR GLN GLN MET PHE ARG PHE PHE GLN GLY VAL GLN LEU ALA PHE
T5
T6
Th4
Th5
61 70 80
LEU VAL GLN SER PRO LYS PHE ALA ALA HIS THR GLN ARG VAL VAL SER ALA LEU ASP GLN
T7
T8
Th6
V1
C1
81 90 100
THE LEU LEU ALA LEU ASN ARG PRO SER ASP GLN PHE VAL TYR MET ILE LYS GLU LEU GLY
T8
Th7
Th8
Th10
Th11
Th12
Th9
101 110 120
LEU ASP HIS ILE ASN ARG GLY THR ASP ARG SER PHE VAL GLU TYR LEU LYS GLU SER LEU
T9
T10
Th13
Th14
V2
V3
C2
121 130 140
GLY ASP SER VAL ASP GLU PHE TRE VAL GLN SER PHE GLY GLU VAL ILE VAL ASN PHE LEU
Th15
V4
V5
C2
141 147
ASN GLU GLY LEU ARG GLN ALA
C3

Fig. 5 : Summary of the data used to establish the complete amino acid sequence of Artemia domain E₁.
T, C, V and Th tryptic, chymotryptic, S. aureus V8 protease and thermolytic peptides respectively.
Automated sequence determination
Manual sequence determination
Identification of the DABTH amino acid residues by
TLC
HPLC
TLC + HPC
Substracted from amino acid composition of peptides

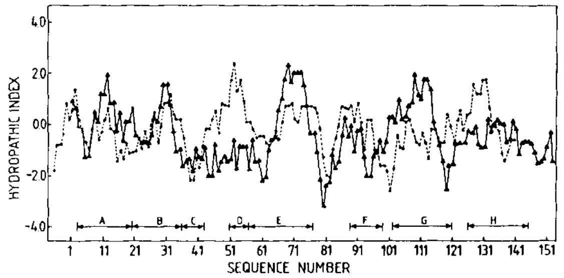


Fig. 11 : Comparison of the hydrophobicity profiles of Artemia domain E₁ and sperm whale myoglobin.
The hydrophobicity profiles were calculated according to Kyte and Doolittle (1982) using a window length of 7.
Artemia domain E₁ · -
Sperm whale myoglobin ---
The standard numbering system based on myoglobin is used.

```
                       10        20          30           40         50          60           70          80          90          100
                             A        8  12 15           D     6   10          C2 4      CD 4    7    D          E     5 78 11
HB ALPHA              V L   SAADKTNVKAAWSKVG     G   HAGEYGAEALERMFLG   FPTTKTY   FPHF  DLS   H        G   SAQVKAHGKKVGDALTLAVG
HB BETA               VHL   SGEEKAAVLALWDKV          NEEEVGGEALGRLLVV   YPWTQRF   FDSFGDLS   NPGAVMG    NPKVKAHGKKVLHSFGEGVH
MYO WHALE             V L   SEGEWQLVLHVWAKVE     A   DVAG GODILIRLFKS   HPETLEK   FDRFKHLK   TEAEMKA    SEDLKKHGVTVLTALGAILK
ARTEMIA E1  ERVDPIT Q L   SQLEKNAILQTWGKVR     G   NLQEVGKATFGKLFAA   HPEYQQM   FRFFQDVQ   LA FLVQ    SPKFAAHTQRVVSALDQTLL
CHIRON III            L   SADQISTVQASFDKVK     G       DPVGILYAVFKA   DPSIMAK   FTQFAGKD   LE SIKG    TAPFETHANRIVGFFSKIIG
CHIRON. IX     DP     V   SSDEANAIRASWAGVK           HN EV  DILAAVFSD   HPDIQAR   FPQFAGKD   LA SIKD    TGAFATHAGRIVGFISEIVA
PETROMYZON  PIVDTGSVAPL   SAAEKTKIRSAWAPVY     S   NYETSGVDILVKFFTS   TPAAQEF   FPKFKGLT   TADQLKK    SADVRWHAERIINAVNDAVA
SCAPHARCA   PSVYDAA AQL   TADVKKDLRDSWKVIG     S   DKKGNGVALMITLFAD   NQETIGY   FKRLGNVS   QGMA       NDKLRGHSITLMYALQNFID
APLYSIA               S L   SAAEADLAGKSWAPVF     A   NKNANGADFLVALFEK   FPDSANF   FADFKGKS   VAD IKA    SPKLRDVSSRIFTRLNEFVN
ANADARA     PSVQDAA AQL   TADVKKDLRDSWKVLG     S   DKKGDGMALMTTLFND   HQETIAY   FKRMGDVS   QGMA       NSKLRGHSITLMYALQNFID
GLYCERA               G L   SAAQRQVIAATWKDIA     G NDNGAGVGKDCLIKHLSA   HPQMAAV   F GFSG        AS      DPAVADLGAKVLAQIGVAVS
TYLO 1           TDCGI L   QRIKVKQQWAQVYSV     G   ESRTDFAIDVFNNFFRT   N?D RSL   FNRVNGDN   VYS         PEFKAHMVRVFAGFDILIS
LUMBRICUS       KKQCGV L   EGLKVKSEWGRAYGS     G   HDREAFSQAIWRATFAQ   VPESRSL   FKRVHGDH   TSD         PAFIAHAERVLGGLDIAIS
VITREO                M L   DQQTINIIKATVPVLK     E    HGVTTTTFYKNLFAK   HPEVRPL   FDMGRQES   LEQP KAL   AMTVLAAAQNIENLPAILPA
LEG-HB               GA L   TESQAALVKSSWEEFN     A    NIPKHTHRFFILVLEI   APAAKDL   FS FLK G   TSEVPQN    NPELQAHAGKVFKLVYEAAI
```

```
                  110       120         130         140        150        160           170         180         190         200
            EF                          F   4   8    FG         G                 16      GH   5   H          10              22
HB ALPHA    HL     D  DLPGA             LSNLSDLHAH   KL   RV   DPVNFKLLSHCLLSTLAVH   LPNDF   TPAVHASLDKFL  SS VSTVLTSKYR
HB BETA     HL     D  NLKGT             FAALSELHCD   KL   HV   DPENFRLLGNVLVVVLARH   FGKDF   TPELQASYQKVV  AG VANALAHKYH
MYO WHALE   K      K GHHEAE             LKPLAQSHAT   KH   KI   PIKYLEFISEAIIHVLHSR   HPGDF   GADAQGAMNKAL  ELFRKDIAAKYKELGYQG
ARTEMIA E1  AL     N  RPSDQFVYH         IKELGLDHIN    R   QT   DRSFVEYLKESLGDSVDEF   TVQSF   GEVIVNFLNEGL  RQA
CHIRON III  EL     P  NIEAD             VNTFVASHKP    R   GV   THDQLNNFRAGFVSYMKAH   T   DF   A GAEAAWGATLDTFF  GMIFSKM
CHIRON. IX  LV     G  NESNAPAMATL       INELSTSHHN    R   GI   TKGQFNEFRSSLVSYLSSH   A   SW   NDATADAWTHGLDNIF  GMIFAHL
PETROMYZON  SM    DDTEKMS               MKNLSGKHAK   SF   QV   DPQYFKVLAAVIADTVAAG                DAGFEKLM  SMICILLRSAY
SCAPHARCA   QLQNPDDLVCV                 VEKFAVNHIT    R   KI   SAAEFGKINGPIKKVLASK   NFGDK   YANAWAKLVAVVQAAL
APLYSIA     DAANAGKMSAM                 LSQFAKEHVG   FG    V   GSAQFENVRSMFPGFVASV   AAP      PAGADAWTK L  FGLIIDALKAAGK
ANADARA     QLDSTDDLICV                 VEKFAVNHIT    R   KI   SGAEFGKINGPMKKVLASK   NFGDK   YANAWAKLVGVVQAAL
GLYCERA     HL   GDZGKMVAQ              MKAVGVRHKG   YGNKHI    KGQYFEPLGASLLSAMEHR   IGGKM   NAAAKDAWAAAY  AD ISGALISGLQS
TYLO 1      VL     D  DKFVLDQA          LAHYAAFHKQ    F  GTI   PFKAFGQTMFQTIAEHIHG   AD                     IGAWRACYAEQIVTGITA
LUMBRICUS   TL     D  QPATLKEE          LDHLQVQHEG    R   KI   PDNYFDAFKTAILHVVAAQ   LGERC    YSNNEEIHDAIACDGFARVLPQVLERGIKGHH
VITREO      VK     K  IAVKHC            QAGVAAAHYP    I    V   GQELLGAIKEVLGDAATDDI  L DAW   G KAYGVIADVF  IQ VEADLYAQAVE
LEG-HB      QLEVTGVVASDAT               LKNLGSVHVS   KG    V   ADAHFPVVKEAILKTIKEV   VGAKW   SEELNSAWTIAY  DELAI VIKKEMDDAA
```

Fig. 8 : Alignment of some selected globin sequences with Artemia domain E₁. The sequences as given in the NBRF databank were used. The standard numbering system based on myoglobin is used.